The impact of incomplete data on quantile regression for longitudinal data

Anneleen Verhasselt, **Alvaro J. Flórez**, Geert Molenberghs and Ingrid Van Keilegom

July 17, 2020
Workshop on the Art of Learning with Missing Values (Artemiss)

UHASSELT
I-BioStat
KU LEUVEN
Interuniversity Institute for Biostatistics
and statistical Bioinformatics

Model and univariate quantile regression

$\mathbf{Y}_i = (Y_1, \ldots, Y_n)'$ is an n-dimensional response vector for individual $i = 1, \ldots, N$. Consider the multivariate regression model:

$$\mathbf{Y}_i = \mathbf{X}_i \boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

where $\mathbf{X}_i$ is a $(n \times p)$-design matrix of covariates, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$ is a vector of regression coefficients, and $\boldsymbol{\varepsilon} = (\varepsilon_{i1}, \ldots, \varepsilon_{in})$ is a vector of error terms.

Then, assuming that $Q_\tau(\boldsymbol{\varepsilon}_i | \mathbf{X}_i) = \mathbf{0}$, the τ-th conditional quantile of $\mathbf{Y}_i$ is:

$$Q_\tau(\mathbf{Y}_i | \mathbf{X}_i) = \mathbf{X}_i' \boldsymbol{\beta}.$$

Univariate quantile regression (UQR):

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{N} \sum_{j=1}^{n} \rho_\tau(Y_{ij} - \mathbf{x}_{ij}' \boldsymbol{\beta}),$$

where $\mathbf{x}_{ij}$ is the jth row of $\mathbf{X}_i$, $\rho_\tau(u) = u[\tau - I(u < 0)]$ is the check-loss function used in quantile regression.

Multivariate quantile regression

We propose a **maximum likelihood estimator (MLE)** based on the use of multivariate AL distribution, with density:

$$f_{\mathbf{Y}}(\mathbf{y};\boldsymbol{\theta}) = \frac{2\exp\left[(\mathbf{y}-\mathbf{X}_i\boldsymbol{\beta})'\boldsymbol{\Delta}^{-1}\boldsymbol{\Sigma}^{-1}\boldsymbol{\xi}\right]}{(2\pi)^{n/2}|\boldsymbol{\Delta}\boldsymbol{\Sigma}\boldsymbol{\Delta}|^{1/2}}\left(\frac{(\mathbf{y}-\mathbf{X}_i\boldsymbol{\beta})'(\boldsymbol{\Delta}\boldsymbol{\Sigma}\boldsymbol{\Delta})^{-1}(\mathbf{y}-\mathbf{X}_i\boldsymbol{\beta})}{2+\boldsymbol{\xi}'\boldsymbol{\Sigma}\boldsymbol{\xi}}\right)^{\nu/2}\times$$
$$\times K_\nu\left[\sqrt{(2+\boldsymbol{\xi}'\boldsymbol{\Sigma}\boldsymbol{\xi})(\mathbf{y}-\mathbf{X}_i\boldsymbol{\beta})'(\boldsymbol{\Delta}\boldsymbol{\Sigma}\boldsymbol{\Delta})^{-1}(\mathbf{y}-\mathbf{X}_i\boldsymbol{\beta})}\right],$$

where $\boldsymbol{\Delta} = \text{diag}(\delta_1,\ldots,\delta_n)$, $\delta_j > 0$ (for $j = 1,\ldots,n$), $\boldsymbol{\xi} = (\xi_1,\ldots,\xi_n)'$, $\xi_j = \frac{1-2\tau}{\tau(1-\tau)}$ for $j = 1,\ldots,n$, $\boldsymbol{\Lambda} = \text{diag}(\lambda_1,\ldots,\lambda_n)$, $\lambda_j^2 = \frac{2}{\tau(1-\tau)}$, and $\boldsymbol{\Psi}$ is a correlation matrix.

Alternatively, we consider a **pairwise estimator (PWE)** which maximizes:

$$p\ell(\boldsymbol{\theta}) = \sum_{i=1}^{N}\sum_{s\in S}\varphi_s \log f_{\mathbf{Y}^{(s)}}(\mathbf{y}_i^{(s)};\boldsymbol{\theta}^{(s)}),$$

where $\varphi = \{\varphi_s | s \in S\}$, S is the set of all vectors of length n consisting of zeros and ones, with each vector having exactly two non-zero entries, and $\mathbf{Y}_i^{(s)}$ the subvector of $\mathbf{Y}_i$ corresponding to the components of s that are non-zero.

Quantile regression with missing data

For non-fully-likelihood-based methods (UQR and PWE), we contemplate **inverse probability weighting (IPW)** methods.

For UQR, the IPW estimator of $\boldsymbol{\beta}$ is:

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \sum_{j=1}^{n_i} \frac{R_{ij}}{\pi_{ij}} \rho_\tau(Y_{ij} - \mathbf{X}'_{ij}\boldsymbol{\beta}),$$

For the PWE, we maximize following weighted pseudo-likelihood function:

$$p\ell(\boldsymbol{\theta}) = \sum_{i=1}^{N} \sum_{s \in S} \frac{R_i^{(s)}}{\pi_i^{(s)}} \log f_{\mathbf{Y}^{(s)}}(\mathbf{y}_i^{(s)}; \boldsymbol{\theta}^{(s)}),$$

The probabilities π_{ij} $(j = 2, \ldots, n_i)$ are obtained as follows (assuming that the first time point is always observed):

$$\pi_{ij} = p_{ij} \prod_{l=2}^{j-1} (1 - p_{il}), \text{ if the subject drops out at occasion } j,$$

with p_{il} as the probability of dropping out at occasion l given the subject is still in the study. In practice, p_{il} is unknown and need to be estimated, e.g., using logistic regression model.

Simulation results and final remarks

Based on a simulation with $n = 2$:

Regarding longitudinal data:
The estimators based on the multivariate AL distribution (**MLE** and **PWE**) take into account the dependence structured of the data, and therefore, are more efficient than the UQR. However, they computationally more intensive.

Regarding missing data:
Since the **UQR** and **PWE** are non-likelihood-based method, the analysis of the "complete cases" provide biased estimates. The **IPW** approach successfully correct the bias. However, there is a cost in the efficiency.

Further work:
Consider an **augmented inverse probability weighting (AIPW)** approach to improve efficiency.

Evaluate the estimators for high-dimensional data with a wide range of dependence structures.